Northumbria University
NEWCASTLE

University**Library**

Genomic epidemiology of SARS-CoV-2 in a UK university identifies dynamics of transmission

Dinesh Aggarwal (dinesh.aggarwal@nhs.net)
University of Cambridge https://orcid.org/0000-0002-5938-8172

Ben Warne
University of Cambridge https://orcid.org/0000-0003-1326-0373

Aminu S. Jahun
University of Cambridge, Department of Pathology, Division of Virology, Cambridge, UK

William Hamilton
University of Cambridge, Department of Medicine, Cambridge, UK

Thomas Fieldman
University of Cambridge, Department of Medicine, Cambridge, UK

Louis du Plessis
Department of Zoology, University of Oxford, UK

Verity Hill
5. Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, Scotland
https://orcid.org/0000-0002-3509-8146

Beth Blane
University of Cambridge https://orcid.org/0000-0002-0996-9091

Emmeline Watkins
Public Health Directorate, Cambridgeshire County Council and Peterborough City Council, UK

Elizabeth Wright
Public Health Directorate, Cambridgeshire County Council and Peterborough City Council, UK

Grant Hall
University of Cambridge, Department of Pathology, Division of Virology, Cambridge, UK

Catherine Ludden
University of Cambridge, Department of Medicine, Cambridge, UK

Richard Myers
Public Health England, 61 Colindale Ave, London, NW9 5EQ, UK

Myra Hosmillo
University of Cambridge

Yasmin Goodfellow
University of Cambridge, Department of Pathology, Division of Virology, Cambridge, UK

Malte L. Pinckert
University of Cambridge, Department of Pathology, Division of Virology, Cambridge, UK

Iliana Georgana
University of Cambridge, Department of Pathology, Division of Virology, Cambridge, UK
Rhys Izuagbe
University of Cambridge, Department of Pathology, Division of Virology, Cambridge, UK
https://orcid.org/0000-0001-8080-2510
Danielle Leek
University of Cambridge, Department of Medicine, Cambridge, UK
Olisaeloka Nsonwu
Public Health England, 61 Colindale Ave, London, NW9 5EQ, UK
Gareth Hughes
Public Health England, 61 Colindale Ave, London, NW9 5EQ, UK
Simon Packer
Public Health England, 61 Colindale Ave, London, NW9 5EQ, UK https://orcid.org/0000-0002-9157-3650
Andrew J. Page
Quadram Institute Bioscience, Norwich Research Park, Norwich, UK
Marina Metaxaki
University of Cambridge, Department of Medicine, Cambridge, UK
Stewart Fuller
University of Cambridge, Department of Medicine, Cambridge, UK
Gillian Weale
Health, Safety & Regulated Facilities Division, University of Cambridge, Cambridge, UK
Jon Holgate
University Information Services, University of Cambridge, Cambridge, UK
Christopher A. Brown
Cambridge Covid-19 Testing Centre, Discovery Sciences, R&D, AstraZenenca, Cambridge, UK
The Cambridge Covid-19 testing Centre
University of Cambridge (see Supplementary File for full list of authors)
University of Cambridge Asymptomatic COVID-19 Screening Programme Consortium
University of Cambridge (see Supplementary File for full list of authors)
The COVID-19 Genomics UK (COG-UK) Consortium
Ministerial Correspondence and Public Enquiries Unit, Department of Health and Social Care (see Supplementary File for full list of authors)
Rob Howes
Cambridge Covid-19 Testing Centre, Discovery Sciences, R&D, AstraZenenca, Cambridge, UK
Duncan McFarlane
Institute for Manufacturing, University of Cambridge, Cambridge, UK
Gordon Dougan
University of Cambridge, Department of Medicine, Cambridge, UK

Oliver G. Pybus
Department of Zoology https://orcid.org/0000-0002-8797-2667
Daniela De Angelis
University of Cambridge
Patrick H. Maxwell
University of Cambridge
Sharon J. Peacock
University of Cambridge https://orcid.org/0000-0002-1718-2782
Michael Weekes
University of Cambridge https://orcid.org/0000-0003-3196-5545
Chris Illingworth
University of Cambridge
Ewan M. Harrison
Wellcome Sanger Institute, Hinxton, Cambridge, UK
Nicholas J. Matheson
University of Cambridge
Ian G. Goodfellow
University of Cambridge

Article

Keywords: Higher Education Settings, Asymptomatic Student Screening, Symptomatic Testing, Phylogenetic Comparison, Onward Transmission

DOI: https://doi.org/10.21203/rs.3.rs-520627/v1

Abstract

Understanding the drivers for spread of SARS-CoV-2 in higher education settings is important to limit transmission between students, and onward spread into at-risk populations. In this study, we prospectively sequenced 482 SARS-CoV-2 isolates derived from asymptomatic student screening and symptomatic testing of students and staff at the University of Cambridge from 5 October to 6 December 2020. We performed a detailed phylogenetic comparison with 972 isolates from the surrounding community, complemented with epidemiological and contact tracing data, to determine transmission dynamics. After a limited number of viral introductions into the university, the majority of student cases were linked to a single genetic cluster, likely dispersed across the university following social gatherings at a venue outside the university. We identified considerable onward transmission associated with student accommodation and courses; this was effectively contained using local infection control measures and dramatically reduced following a national lockdown. We observed that transmission clusters were largely segregated within the university or within the community. This study highlights key determinants of SARS-CoV-2 transmission and effective interventions in a higher education setting that will inform public health policy during pandemics.

Main

The SARS-CoV-2 pandemic has caused substantial morbidity and mortality globally[1,2]. Universities have been considered conduits for transmission due to extensive social networks of young adults, many of whom live communally, and in-person teaching of large groups[3]. Outbreaks of SARS-CoV-2 have been observed in a number of higher education institutions, but the drivers for transmission in these settings are poorly understood[4]. It is speculated that infection dynamics are dependent on transmission chains involving student courses, residence, study year and social networks[5]. Understanding these dynamics is essential in order to devise effective infection control measures while minimising disruption to teaching, research and the mental health of students and staff[6]. Furthermore, while university students are less likely to develop severe illness secondary to SARS-CoV-2, there is concern that university outbreaks could seed infections in more vulnerable populations, including staff, the local community, and upon returning home to older relatives[7]. Identifying possible sources of cross-transmission is therefore vital.

Although SARS-CoV-2 genome sequencing has clear utility to identify virus emergence and cryptic transmission[8,9], no large-scale genomic studies in university settings have been conducted. The United Kingdom has an extensive community genomics surveillance programme through COG-UK[10] which complements traditional contact tracing approaches by providing understanding of circulating viral populations.

We report the results of a genomic epidemiology study of SARS-CoV-2 across a complete term at the University of Cambridge (UoC). From 5th October to 6th December 2020, the UoC ran PCR-based symptomatic testing for all staff and students, and offered asymptomatic screening to 15,500 students

living in university-managed accommodation. We therefore provide a unique study of SARS-CoV-2 infection that encompasses pre-symptomatic and asymptomatic students[11]. Positive samples from the UoC were sequenced and compared with systematic surveillance SARS-CoV-2 sequences from the local community. The results were analysed in conjunction with epidemiological data derived from the screening programme and national contact tracing. Overall, we describe introductions of SARS-CoV-2 into a higher education setting, the dynamics of transmission both within the university and between the university and the surrounding community, and the impact of local and national measures to control the spread of SARS-CoV-2 infections.

Results

In total, 972 SARS-CoV-2 cases were identified among university students and staff over the course of term (5th October to 6th December 2020). High-quality genomes were recovered from 446/778 (57.3%) positive cases from the university testing programme. High-quality genomes were recovered from 107/266 (40.2%) cases identified through the Healthcare worker (HCW) screening programme (95 HCWs, 8 students, 4 university staff) and 104 patients identified by hospital testing (71 SARS-CoV-2 positive patients from Cambridge University Hospitals (CUH) and 33 from other medical facilities in Cambridgeshire). A further 797 local cases identified by community testing during the study period were present within the COG-UK dataset, of which 17 were identified as students, 7 as university staff and 26 as HCWs (**Figure 1**). Of all identified SARS-CoV-2 cases from Cambridgeshire (university and community) during this period, 8.0% were sequenced (**Extended Data Figure 1**).

SARS-CoV-2 lineages and transmission clusters

Over the 9-week term, 62 Pango lineages were identified across the university and community (**Figures 2a and 2c**). In the university, 23 Pango lineages were identified, and 438/482 (90.9%) cases were from just 4 lineages (B.1.60.7, B.1.177, B.1.36, B.1.177.16), all of which were detected by the second week of term. 12 lineages were only observed after the second week of term and accounted for 6.9% cases. By comparison, 57 lineages were identified in the local community over the same 9-week period. Viral genomes containing mutations in the spike protein that have been linked to decreased sensitivity to antibody-mediated immunity or impact viral transmission were observed in the university population; 3 sequences from the B.1.258 lineage containing the N439K mutation and ΔH69/ΔV70, 2 cases of B.1.1.7 and its associated mutations[12], and 88 cases of B.1.177 with the A222V mutations[13].

In total, 198 putative transmission clusters were defined by CIVET, including 16 clusters of 2 or more university members. Only 8 clusters contained 5 or more university members (range 6-337), which represented 91.3% of all university cases, signifying that the majority of introductions into UoC did not cause ongoing transmission. To further investigate the largest of these, cluster 1 described below, we identified groups of identical samples (0 SNP differences) which produced 19 additional clusters (a total of 34 university clusters) for further analysis.

Determinants of viral spread across the university

To determine transmission dynamics following introduction into the university, we performed a detailed investigation of the largest genomic cluster (Cluster 1), which accounted for 337/484 (69.6%) sequenced university cases (**Figure 3**). This was widely dispersed across the university by the middle of term, affecting students from 29/31 Colleges, 28 undergraduate courses and 208 households in university accommodation alone (**Figure 4**).
Cluster 1 was classified as belonging to Pango lineage B.1.160.7. No mutations previously noted to be associated with increased transmissibility were observed in this lineage compared to other genomes in the study. Interrogation of the entire COG-UK dataset of samples from 2020 showed that this lineage was first identified in the UK on 4th October 2020, in Wales, before becoming predominantly sampled in UoC (**Figure 3b**). The B.1.160.7 lineage was not identified in the local community until term week 3, suggesting that the university cases were introduced from outside Cambridgeshire. This was supported by the median estimate of the time to the most common recent ancestor of cluster 1, in comparison to its most closely related cluster from Cambridgeshire community isolates of 115 days (C.I. 91-148) prior to the start of term. Additional analysis with A2B-COVID showed that these sequences were consistent with a single introduction into the university (**Figure 3c**).

National and university contact tracing data were used to identify the initial source of dispersion of this cluster. Ten students from the first two weeks of term reported visiting the same nightclub (venue A). Nine individuals either had an isolate from cluster 1 or (in the event that their sample did not yield a high-quality sequence) were household contacts of an individual with a sequenced cluster 1 isolate. No information was available for one student.

Transmission of cluster 1 was sustained from the first week of term until a national lockdown was enforced on 5th November. Students testing positive in the two weeks around lockdown reported common exposure events predominantly linked to nightclub venues (25/59 (42.4%) of exposures external to the university reported by 48 students). Venue A, identified above as the possible source of dispersion of this cluster at the start of term, was also the most common venue identified in the two weeks around lockdown (n=16). 9/16 cases had sequences in cluster 1, and a further 5 individuals (where no sequence was available) were household contacts of sequenced cases in cluster 1 (**Extended Data Figure 5**).

To determine the impact of lockdown and other control measures within the university, a birth-death skyline model[14] was used to measure changes in the effective reproduction number (R_e) within cluster 1. The model indicated an initial R_e at the start of term that was slightly larger than 1, albeit with wide uncertainty (median 1.11; 95% HPD: 0.24-2.08 on 5th October). Over the next 2 weeks R_e continued to rise (median 1.54; 95% HPD 1-2.22 on 15th October) followed by a subsequent gradual decline over the next 2 weeks (**Figure 5a**). There was a rise immediately prior to the start of lockdown (median 1.53; 95% HPD 1.24-1.84 on 5th November), followed by a steep decrease thereafter (median 0.25; 95% HPD 0.09-0.44 on 19th November) (**Figure 5a**), consistent with declining absolute numbers of SARS-CoV-2 infections seen

during this time (**Figure 2c**). The model estimated the median effective infectious period for individuals in the cluster at 2.91 days (95% HPD: 2.38-3.47 days) (**Figure 5b**). As the model does not explicitly incorporate an incubation period and assumes that individuals cannot transmit after being sampled, the effective infectious period represents the mean time from infection until testing positive and assumes perfect infection control measures thereafter. Estimates of R_e and the effective infectious period are robust to model parameterisations (**Extended Data Figures 6 and 7**). Sampling proportion estimates largely overlap with empirical estimates based on the number of positive cases that were sequenced during each week (**Figure 5c**). Although sampling proportion estimates are sensitive to the prior specifications, R_e estimates are unaffected (**Extended Data Figure 8**).

Transmission within university households

There was evidence of transmission of SARS-CoV-2 in student accommodation in 18/34 university clusters. In cluster 1, 169/337 (50.1%) students had a virus genome sequence identical to at least one other student living in the same or neighbouring household (sub-clusters within 0 SNPs ranging between 2-11 students).

The largest cluster associated with transmission in accommodation was cluster 2 (lineage B.1.36). By term week 3, this cluster involved 30 students, of which 24 (80%) lived in the same accommodation block in College A and 4 students lived in 2 separate households in the same college (**Extended Data Figure 9**). Interventions from the university, supported by local public health authorities, included isolation of all households in the main accommodation block and individual screening offered to all students. Half of all cases in this cluster were diagnosed by asymptomatic screening. No further genomically-related isolates were identified after term-week 3, indicating a successful intervention, and cessation of transmission.

To quantify the importance of household transmission, a Reed-Frost Chain Binomial Model was employed to estimate the household attack rate. 265 households in which the data were consistent with only 1 introduction of SARS-CoV-2 were identified using A2B-COVID. The per household contact probability that an infected person passed on the virus to an uninfected individual within the same household was estimated at 7.8% (95% C.I. 6.9-8.7%).

Further genomic clusters where transmission between household members was implicated are outlined in supplementary table 1. They follow similar patterns, with groups of cases confined to a single college not leading to sustained transmission.

Other transmission routes among university members

In addition to household transmission, there was evidence of viral spread between students in the same course and year of study in 14/34 genomic clusters, with the highest proportion being students in their first year of study. In cluster 1, 203/337 (60.2%) students had an identical isolate to at least one other

student studying the same course in the same year (cluster size range 2-14 students). Statistical modelling using data from cluster 1 across the term showed a bias towards infections being observed in first year students (*p*-value=0.002) (**Extended Data Figure 10**, model details in supplementary methods). Two further small clusters comprise postgraduate students working in the same university department.

However, we were not able to determine the probable location of transmission in most cases: there is considerable overlap between course and household clusters, as well as complex social and study networks between students (illustrated in supplementary table 1, for example in clusters 3, 4 and 10). Of note, 23/34 clusters with 2 or more genomically linked cases in the dataset contained at least one university member that could not be epidemiologically linked with any other case in their cluster.

The number of SARS-CoV-2 sequences from university staff members were limited in comparison to students (n=30). There was evidence of transmission between staff members working in the same department, college or ancillary role in four genomic clusters. Two clusters contained staff members who shared the same household. There are 8 clusters involving both university staff and students. However, epidemiological associations between these two groups could only be identified in one cluster: a shared household between a student and staff member working in separate university departments.

Transmission between the university and local community

We next sought to address the degree of transmission between the university and the local community. Two distinct phylogenetic approaches, shown in figure 2, demonstrate segregation of the majority of community and university cases into separate clusters and therefore a lack of substantial cross-transmission. Of the 198 clusters across the dataset, 29 (14.6%) contained both university and community cases. Only 6 clusters contained 5 or more university cases and included 3 or more community cases.

CIVET was run separately with university and hospital (patient and healthcare worker) cases for a focused phylogenetic analysis of this setting. Associations were identified between university and hospital settings, with 17 clusters involving both university members and either patients or staff. Cluster 1 (69.6% of student cases), contained only 1 patient and 1 healthcare worker with no identifiable epidemiological link to students. The remaining 16 clusters comprised 133 individuals, including 26 patients, 55 hospital staff or their family members and 52 university members (including 18 staff and 15 clinical medical students). The second largest cluster of university members (n=21 university and hospital cases) included 9 medical students, 5 healthcare workers and 2 patients. Phylogenetically, the medical students and one of the healthcare workers were closely linked (**Extended Data Figure 11**) and analysis of these cases with A2B-COVID confirmed plausible transmission. All 9 medical students were on clinical rotations at the time of diagnosis of the index case; 7/9 lived in neighbouring households in the same college and the remaining 2 were named contacts of the index student. Plausible transmission events between this group and the other cluster members were refuted using A2B-COVID (**Extended Data Figure 11**).

To further investigate epidemiological associations in clusters involving university members and the local community, 1243/1455 of the cases sequenced over the sampling period were linked to national contact tracing data (excluding hospital cases). 219 (17.6%) cases reported 127 common exposure events. Cluster 1, representing 69.6% of cases within the university, included only 18/976 (1.8%) community cases; only one community case had a common exposure with a university student, dining at the same restaurant. No other epidemiological links were identified in all other genomic clusters. Transmission suspected in 19 epidemiologically linked clusters defined by common exposures was refuted by phylogenetic variation.

Discussion

We report the first comprehensive and integrated epidemiological and genomic analysis of SARS-CoV-2 transmission in a higher education setting. Following a limited number of introductions, the majority of cases were linked to a single genetic cluster, that was likely to have dispersed across the university following multiple social gatherings at a nightclub. There was considerable transmission associated with student accommodation and student courses, but minimal evidence of transmission in departments, or between students and staff. We observe the great majority of transmissions occur either within the university or within the local community. Finally, we present evidence demonstrating the efficacy of university measures and national lockdown in reducing COVID-19 cases.

Nearly 70% of all university cases belonged to one genetic cluster (cluster 1), introduced into the UoC by the arrival of students and likely forming a single transmission chain. A nightclub was implicated as an important transmission event at the start of term and again prior to lockdown. This corroborates previous studies identifying such venues as a risk factor for substantial SARS-CoV-2 transmission[15,16]. We urge a cautious approach to access of such venues during a SARS-CoV-2 pandemic, particularly in the context of a young susceptible student population.

Our data showed a significant rise in the effective reproduction number coinciding with the announcement of a national lockdown on 31st October to begin on 5th November 2020. This supports the need for a targeted public health campaign to limit higher risk activities prior to the implementation of socially restrictive measures. National lockdown dramatically reduced transmission, demonstrating both compliance and an effective control strategy.

Multiple findings in this study demonstrate the efficacy of university strategies to control transmission. First, we highlight a limited number of introductions and low lineage diversity in the university compared to the surrounding community. While natural extinction of lineages is relatively common[17], multiple genomically diverse clusters may be expected given the congregation of students from across the globe (international students make up 35% of students in college accommodation[11]. The lack of diversity may reflect the impact of robust and widely implemented university infection control measures, such as social distancing, mask wearing and quarantine of international students at the beginning of term.

Further, although we have demonstrated that transmission between students in the same accommodation block is an important factor in the spread of SARS-CoV-2, we report a lower secondary household attack rate (7.8%) than that identified in domestic households (16.6%-21.1%)[18,19], a lower than expected effective infectious period (2.9 days)[20], and a successful reduction of the effective reproduction number in the university to below 1 after an initial rise at the start of term. In multiple clusters, transmission in student households was successfully interrupted through measures provided by the university, including rapid case identification through asymptomatic screening and the availability of symptomatic testing, contact tracing and comprehensive support provided by colleges for cases and their contacts while in isolation. In identifying considerable transmission between students on the same course, we suggest that further mitigation of viral spread may be obtained by implementing shared student accommodation based on university courses.
Finally, we observed limited transmission between the university and the local community. The largest university cluster, accounting for the majority of student infections, was largely phylogenetically distinct from community cases. Further, epidemiological evidence describing common exposures for community and university cases was sparse. However, clinical medical students were disproportionately represented within community clusters. This is an important epidemiological link between secondary care and the university; we highlight this group as being at-risk for both acquisition and transmission of SARS-CoV-2 and medical students should therefore be prioritised for interventions such as vaccination.

A combination of contact tracing and genomics was instrumental to understanding transmission within the university and with its surrounding population; we advocate for a combined genomic epidemiological approach to inform outbreak investigations.

This study has a number of limitations. Incomplete sampling and subsequent sequence filtering in both the university and community should be considered when interpreting transmission; the asymptomatic and active case ascertainment in this study should mitigate this discrepancy. The lower community case ascertainment may result in unobserved transmission chains. We highlight shared student courses as a risk factor for transmission; this does not take into account the setting of transmission, i.e. during educational or social activities. Finally, the UoC is distinct in its collegiate structure with limited integration with the community; any generalisation of conclusions should be tempered by the study setting.

Conclusion

We present the first comprehensive integrated epidemiological and genomic evaluation of transmission of SARS-CoV-2 within a university. The insights gained will inform public policy regarding infection control measures in higher education settings. We find containment of transmission in student accommodation necessary to mitigate onward propagation. We highlight the importance of targeted public health measures towards nightclub venues to limit transmission. Critically, these findings are likely to be informative for future pandemic preparedness.

Methods

Study Setting

The UoC has approximately 23,000 students and 12,600 staff. The university is divided into 31 colleges and 150 departments, faculties and other institutions. Students belong to a college community, as well as being members of the university and an academic faculty/department. Colleges provide residential accommodation for approximately two thirds of students, either on campuses or in off-site housing, and offer social and sports activities, pastoral and academic support for each individual[21]. All Colleges have membership from students across multiple courses. The university is based in the City of Cambridge (which has an estimated population of 123,900[22]), in the county of Cambridgeshire (estimated population 855,796 people in 2019[23]) in the East of England.

Participants and samples

Samples were derived from university symptomatic testing and asymptomatic COVID-19 screening programmes between October 5 2020 and December 6 2020, covering the full term. Testing for all symptomatic students and staff was available on weekdays. The asymptomatic screening programme has been described in detail elsewhere[11]. In brief, screening was offered on a voluntary basis to all students residing in accommodation owned or managed by a college or the Cambridge Theological Federation. In total, 15,561 students were eligible to participate. To optimise testing efficiency, multiple swabs were pooled into the same tube of viral transport medium at the time of sample collection. Testing pools varied in size from 1 to 10 students, with each devised to include one or more student households as far as possible[11]. In this study, households are defined as individuals who share a kitchen, bathroom and/or lounge facilities. The members of any pool testing positive were re-tested using individual confirmatory PCR tests to confirm the result and identify the positive subject(s) (see supplementary methods for further details including infection prevention control measures). Only samples from individuals that were confirmed positive upon the re-testing were used for sequencing.
SARS-CoV-2 strains circulating in the local community were identified from the COG-UK dataset for Cambridgeshire. These data were derived from local community samples from non-hospitalised, symptomatic individuals, who requested a free diagnostic test via national community testing. Other samples were derived from patients treated at three Cambridgeshire hospital trusts: Cambridge University Hospitals NHS Foundation Trust (a teaching hospital providing secondary care services for Cambridge and the surrounding area as well as tertiary referral services for the East of England and surge capacity for COVID-19); Royal Papworth Hospital NHS Foundation Trust (specialist heart and lung hospital, also providing surge capacity for COVID-19); Cambridgeshire and Peterborough NHS Foundation Trust (provider of community, mental health and learning disability services in Cambridgeshire). Hospital samples were obtained from both asymptomatic screening and those exhibiting COVID-19 symptoms. Finally, samples were derived from the asymptomatic HCW programme at Cambridge University Hospitals[24].

Sequencing

Positive samples from UoC testing with a PCR cycle threshold value ≤33 were selected and sequenced using the GridION platform (Oxford Nanopore). All Cambridgeshire samples sequenced between 24th September and 21st December 2020 were included to overlap with the university term. Samples from the local Cambridgeshire community and hospital cases (described above) were collected as part of national SARS-CoV-2 testing, and sequenced at one of seventeen COG-UK sequencing sites (further details in supplementary methods). The samples were prepared using either the ARTIC[25] or veSeq[26] protocols, and were sequenced using Illumina or Oxford Nanopore platforms. Genomic data were filtered to exclude sequences with >5% 'N's and those of spuriously low file sizes (<29KB). Genomes were aligned with minimap2[27] to the Wuhan Hu-1 reference genome (MN908947.3), collected December 2019. All samples were processed through COVID-CLIMB pipelines[28,29]. Protocols are available at https://github.com/COG-UK.

Phylogenetic analysis

Maximum likelihood phylogenetic trees were estimated using IQ-TREE (version 2.1.2 COVID-edition)[30] and rooted using Wuhan Hu-1 (MN908947.3) as an outgroup. The tree was constructed using the GTR+ Γ substitution model[31], as determined by ModelFinder[32]. Branch support statistics were generated using the ultrafast bootstrap method[33]. TempEst[34] was used to explore the temporal signal in the data. Trees were visualised, explored, and labelled with associated metadata using Microreact[35] to identify epidemiological links supported by the genomic data. Specified mutations were identified using type_variants (https://github.com/cov-ert/type_variants). Possible transmission clusters were defined by extracting phylogenetic neighbourhoods identified using the CIVET tool (version 2.1.0) on 2021-01-11 (https://github.com/artic-network/civet). In selected clusters, further evaluation was conducted using A2B-COVID[36]. Where indicated, collapsed nodes from trees generated from CIVET were inspected to visualise data in the context of the COG-UK national database (https://www.cogconsortium.uk/) For further evaluation of transmission in the largest cluster identified by CIVET, pairwise SNP differences between sequences were determined using SNP-dist (version0.7.0) (https://github.com/tseemann/snp-dists).

Lineages

Global Pango Lineages[37] were assigned to each genome using Pangolin v2.1.6 (https://github.com/cov-lineages/pangolin) with analyses performed on COVID-CLIMB[29] (further details in supplementary methods).

Molecular clock and phylodynamic analyses

BEAST v1.10.4[38] was used to perform a time-scaled phylogenetic analysis using an exponential growth coalescent treeprior and an HKY+ Γ substitution model including all university and community high-quality genomes from the study period. As there was a lack of clear temporal signal in our dataset due to the relatively short time period analysed, the substitution rate was fixed to $1x10^{-3}$ substitutions per site per year (s/s/y) under a strict clock model[39] Two chains of 50 million iterations were run independently to ensure convergence to the correct posterior distribution. Convergence was assessed using Tracer[40], and 10% of states were removed to account for burn-in. Finally, a maximum clade credibility (MCC) tree was generated using TreeAnnotator.

To estimate the effective reproduction number (R_e) and infectious period of SARS-CoV-2 over the term, a dominant clade (representing 69.6% of all university genomes) was selected and all community genome sequences that cluster with it incorporated, resulting in a total of 354 genomes. A Bayesian birth-death skyline (BDSKY) model[14] was employed using BEAST v2.6 [41]. An HKY substitution model was used along with a strict clock model, placing a lognormal prior with mean $1x10^{-3}$ s/s/y (in real space) and standard deviation 0.1 on the clock rate. A lognormal prior with mean 0 and standard deviation 1 was placed on R_e and a Beta prior with α=5 and β=5 was placed on the sampling proportion. R_e was parameterised into 20 epochs, equidistantly spaced between the origin time and the most recent sequence collection date. The sampling proportion was fixed to 0 before the first week of term and estimated for each week thereafter. The rate at which infected patients become non-infectious was assumed to be constant and a lognormal prior with mean 48.7 $years^{-1}$ (in real space) and standard deviation 0.25 was placed on it, resulting in a prior mean effective infectious period between ~5 and ~15 days. To test the robustness of the posterior estimates different parameterisations were used for R_e and the sampling proportion, and the sampling proportion prior was varied. Further details are provided in the supplementary methods. For all models two chains of 500 million iterations were run independently. Convergence was assessed using the R-package *coda* [42], and 10% of states were removed to account for burn-in. MCC trees were generated using TreeAnnotator.

Household attack rates

A2B-COVID was used to exclude households for which the sequence and epidemiological data were inconsistent with a single viral introduction to the household. A chain binomial model was then used to estimate the probability that an infected person transmitted the virus to an uninfected person within the same household (further details in supplementary methods).

Epidemiological data

University student demographic data were derived from the UoC student electronic record system CamSIS, and household structure and membership data from the UoC asymptomatic screening programme. To identify university affiliated cases (students and staff) and hospital staff accessing the

national SARS-CoV-2 testing service, Second Generation Surveillance System (SGSS) and contact-tracing data provided by NHS Test and Trace (T&T) data were interrogated. Epidemiologically linked ‘common exposures’ for students, university staff and the local community were identified through T&T data. Common exposures were defined by T&T as locations or events that two or more people testing positive for COVID-19 visited in the same two to seven day period before symptom onset or positive test. Additional contact tracing information was also provided by the UoC COVID helpdesk. These data were compared with observed phylogenetic clusters to determine potential sources of transmission and determine the extent of transmission between the university and community.

Epidemiological data from UoC were initially compiled in Microsoft Azure SQL and Excel 2013 (Microsoft) and analysed in STATA 14.2 (College Station, TX, USA).

Declarations

Funding

DA is a Wellcome Clinical PhD Fellow and gratefully supported by the Wellcome Trust (Grant number: 222903/Z/21/Z). BW receives funding from the University of Cambridge and the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre (BRC) at the Cambridge University Hospitals NHS Foundation Trust. IG is a Wellcome Senior Fellow and is supported by the Wellcome Trust (Grant number: 207498/Z/17/Z and 206298/B/17/Z). EMH is supported by a UK Research and Innovation (UKRI) Fellowship: MR/S00291X/1. CJRI acknowledges Medical Research Council (MRC) funding (ref: MC_UU_00002/11). NJM is supported by the MRC (CSF MR/P008801/1) and NHSBT (WPA15-02). AJP gratefully acknowledge the support of the Biotechnology and Biological Sciences Research Council (BBSRC); their research was funded by the BBSRC Institute Strategic Programme Microbes in the Food Chain BB/R012504/1 and its constituent project BBS/E/F/000PR10352, also Quadram Institute Bioscience BBSRC funded Core Capability Grant (project number BB/CCG1860/1). LdP and OGP were supported by the Oxford Martin School. This research was supported by the NIHR Cambridge BRC. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care. The COVID-19 Genomics UK Consortium is supported by funding from the MRC part of UK Research & Innovation (UKRI), the National Institute of Health Research and Genome Research Limited, operating as the Wellcome Sanger Institute. The Cambridge Covid-19 testing Centre is funded by the Department of Health and Social Care, UK Government. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Author contributions

All authors read the manuscript and consented to its publication.

First	Last Name	Contribution
Dinesh	Aggarwal	a, b, c, d, e, f, h, i, j, k, m
Beth	Blane	f, g, h, j, m
Christopher	Brown	f, g, h
Daniela	De Angelis	j, l
Gordon	Dougan	l
Louis	du Plessis	b, c, e, g, h, j, k
Tom	Fieldman	a, b, d, e, f, h, j
Stewart	Fuller	f, g, h
Iliana	Georgana	e, f, h
Yasmin	Goodfellow	e, f, h
Ian	Goodfellow	a, b, f, g, j, l, m, n
Grant	Hall	e, f, h
Will	Hamilton	a, b, c, e, j, k
Ewan	Harrison	a, b, g, j, l, m, n
Verity	Hill	b, c, e, j
Jon	Holgate	f, g, h
Myra	Hosmillo	e, f, h
Rob	Howes	b, d, g, h
Gareth	Hughes	f, h
Rhys	Izuagbe	e, f, h
Aminu	Jahun	b, e, f, h, j
Danielle	Leek	f
Chris	Illingworth	b, c, e, h, j, k, l
Catherine	Ludden	a, g, j, n
Duncan	McFarlane	g, l, m
Nick	Matheson	a, b, j, l, m, n
Patrick	Maxwell	j, l, m, n
Marina	Metaxaki	f, g, h

Richard	Myers	j, l, m
Olisaeloka	Nsonwu	f, h
Simon	Packer	b, d, g, h
Andrew J	Page	l
Sharon J	Peacock	a, b, g, j, l, m, n
Malte L	Pinckert	e, f, h
Oliver	Pybus	j, l
Ben	Warne	a, b, d, e, f, h, j, k, m
Emmeline	Watkins	a, j
Gillian	Weale	f, g, h
Michael	Weekes	j, m, n
Elizabeth	Wright	f, h, j
Cambridge Covid-19 Testing Centre	-	b, d, g, h
a, conceptualization; b, methodology; c, software; d, validation; e, formal analysis; f, investigation; g, resources; h, data curation; i, writing – original draft preparation; j, writing – review and editing; k, visualization; l, supervision; m, project administration; n, funding acquisition		

Conflicts of interests and disclosures

RH is an employee of AstraZeneca AB.

Ethics

This study was conducted as part of surveillance for COVID-19 infections under the auspices of Section 251 of the NHS Act 2006 and/or Regulation 3 of The Health Service (Control of Patient Information) Regulations 2002. They therefore did not require individual patient consent or ethical approval. Public Health England affiliated authors had access to identifiable Cambridgeshire community case data. Other authors only had access to anonymised or summarised data. The COG-UK study protocol was approved by the Public Health England Research Ethics Governance Group (reference: R&D NR0195). Ethical approval for the University of Cambridge asymptomatic COVID-19 screening programme was granted by the University of Cambridge Human Biology Research Ethics Committee (HBREC.2020.35).

Acknowledgements

We thank members of the COVID-19 Genomics Consortium UK and NHS Test and Trace contact tracers for their contributions to generating data used in this study. We are grateful to all students and staff at the University of Cambridge who have contributed to the COVID-19 response during Michaelmas Term. We are grateful to all staff members of the Cambridge COVID-19 Testing Centre for generating qPCR data.

Data availability

Assembled/consensus genomes are available from GISAID[43] subject to minimum quality control criteria. Raw reads are available from European Nucleotide Archive (ENA) [44]. All genomes, phylogenetic trees, basic metadata are available from the COG-UK consortium website (https://www.cogconsortium.uk/data). For confidentiality reasons, extended metadata[45] (Griffiths et al., 2020) is not publicly available, however some may be available upon request from Public Health England and the University of Cambridge, facilitated by authors.

References

1. Zhou, F*., et al.* Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* **395**, 1054-1062 (2020).
2. Yang, J*., et al.* Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: a systematic review and meta-analysis. *Int J Infect Dis* **94**, 91-95 (2020).
3. Amoolya Vusirikala, H.W., Samuel Jones, Elise Tessier, Ray Borrow, Ezra Linley, Katja Hoschler, Frances Baawuah, Shazaad Ahmad, Nick Andrews, Mary Ramsay, Shamez N Ladhani, Kevin E Brown, Gayatri Amirthalingam. Seroprevalence of SARS-CoV-2 Antibodies in University Students: cross-sectional study, December 2020, England. (2021).
4. Yamey, G. & Walensky, R.P. Covid-19: re-opening universities is high risk. *BMJ* **370**, m3365 (2020).
5. Group, C.s.T.F. Risks associated with the reopening of education settings in September. Vol. 2021 (2020).
6. Sahu, P. Closure of Universities Due to Coronavirus Disease 2019 (COVID-19): Impact on Education and Mental Health of Students and Academic Staff. *Cureus* **12**, e7541 (2020).
7. Education, T.a.F.G.o.H.E.F. Principles for managing SARS-CoV-2 transmission associated with higher education - 3 September 2020. Vol. 2021 (2020).
8. Meredith, L.W*., et al.* Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect Dis* **20**, 1263-1271 (2020).
9. Moreno, G.K*., et al.* SARS-CoV-2 transmission in intercollegiate athletics not fully mitigated with daily antigen testing. *medRxiv* (2021).

10. consortiumcontact@cogconsortium.uk, C.-G.U. An integrated national scale SARS-CoV-2 genomic surveillance network. *Lancet Microbe* **1**, e99-e100 (2020).
11. Ben Warne, J.E., Marina Metaxaki, Stewart Fuller, Richard J Samworth, Gillian Weale, Jon Holgate, Vijay Samtani, Craig Brierley, Dinesh Aggarwal, Sarah Hilborne, Mahin Bagheri Kahkeshi, Sarah Berry, Julie Douthwaite, Alexandra L. Orton, Rebecca Clarke, Darius Danaei, Rory Dyer, Rob Glew, Oliver Lambson, Aastha Dahal, Ben Margolis, Edna Murphy, Matthew Russell, Vickie Braithwaite, Kathryn Faulkner, Elizabeth Wright, Afzal Chaudhry, Lenette Mactavous, Ajith Parlikad, University of Cambridge Asymptomatic COVID-19 Screening Programme Consortium, Jane Greatorex, Ian Leslie, Andy Neely, Steve Rees, Ashley Shaw, Martin Vinnell, Linda Sheridan, Emmeline Watkins, Stephen Baker, Ian G. Goodfellow, Paul J. Lehner, Kathleen Liddell, Rob Howes, Michael P. Weekes, Duncan McFarlane, Patrick H. Maxwell, Nicholas J. Matheson. Feasibility and efficacy of mass testing for SARS-CoV-2 in a UK university using swab pooling and PCR. *bioRxiv* (2021).
12. Andrew Rambaut, N.L., Oliver Pybus, Wendy Barclay, Jeff Barrett, Alesandro Carabelli, Tom Connor, Tom Peacock, David L Robertson, Erik Volz, on behalf of COVID-19 Genomics Consortium UK (CoG-UK). Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. Vol. 2021 (2020).
13. Hodcroft, E.B.*, et al.* Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *medRxiv* (2020).
14. Stadler, T., Kuhnert, D., Bonhoeffer, S. & Drummond, A.J. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc Natl Acad Sci U S A* **110**, 228-233 (2013).
15. Muller, N.*, et al.* Severe Acute Respiratory Syndrome Coronavirus 2 Outbreak Related to a Nightclub, Germany, 2020. *Emerg Infect Dis* **27**, 645-648 (2020).
16. Choi, H., Cho, W., Kim, M.H. & Hur, J.Y. Public Health Emergency and Crisis Management: Case Study of SARS-CoV-2 Outbreak. *Int J Environ Res Public Health* **17**(2020).
17. du Plessis, L.*, et al.* Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371**, 708-712 (2021).
18. Thompson, H.A.*, et al.* SARS-CoV-2 setting-specific transmission rates: a systematic review and meta-analysis. *Clin Infect Dis* (2021).
19. Madewell, Z.J., Yang, Y., Longini, I.M., Jr., Halloran, M.E. & Dean, N.E. Household Transmission of SARS-CoV-2: A Systematic Review and Meta-analysis. *JAMA Netw Open* **3**, e2031756 (2020).
20. Li, R.*, et al.* Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* **368**, 489-493 (2020).
21. Cambridge, U.o. How the University and Colleges work. Vol. 2021.
22. Statistics, O.f.N. 2011 ONS Census. Vol. 2021.
23. Statistics, O.f.N. Cambridgeshire and Peterborough Population Overview Report. Vol. 2021.
24. Rivett, L.*, et al.* Screening of healthcare workers for SARS-CoV-2 highlights the role of asymptomatic carriage in COVID-19 transmission. *Elife* **9**(2020).

25. Quick, J. nCoV-2019 sequencing protocol v3 (LoCost) V.3. Vol. 2021 (2020).
26. Bonsall, D.*, et al.* A Comprehensive Genomics Solution for HIV Surveillance and Clinical Monitoring in Low-Income Settings. *J Clin Microbiol* **58**(2020).
27. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
28. Nicholls, S.M.*, et al.* MAJORA: Continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance. *bioRxiv* (2020).
29. Connor, T.R.*, et al.* CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. *Microb Genom* **2**, e000086 (2016).
30. Minh, B.Q.*, et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534 (2020).
31. Tavare, S. Some probabilistic and statistical problems in the analysis of DNA sequences. (1986).
32. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A. & Jermiin, L.S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**, 587-589 (2017).
33. Minh, B.Q., Nguyen, M.A. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol* **30**, 1188-1195 (2013).
34. Rambaut, A., Lam, T.T., Max Carvalho, L. & Pybus, O.G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* **2**, vew007 (2016).
35. Argimon, S.*, et al.* Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genom* **2**, e000093 (2016).
36. Illingworth, C.J.R.*, et al.* A2B-COVID: A method for evaluating potential SARS-CoV-2 transmission events. *medRxiv* (2020).
37. Rambaut, A.*, et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* **5**, 1403-1407 (2020).
38. Suchard, M.A.*, et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* **4**, vey016 (2018).
39. Rambaut, A. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* **16**, 395-399 (2000).
40. Rambaut, A., Drummond, A.J., Xie, D., Baele, G. & Suchard, M.A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst Biol* **67**, 901-904 (2018).
41. Bouckaert, R.*, et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* **15**, e1006650 (2019).
42. Plummer, M., Best, Nicky, Cowles, Kate, Vines, Karen. CODA: convergence diagnosis and output analysis for MCMC. *R News* **6(1)**, 7–11 (2006).
43. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* **22**(2017).
44. Toribio, A.L.*, et al.* European Nucleotide Archive in 2016. *Nucleic Acids Res* **45**, D32-D36 (2017).

45. Griffiths, E.J.T., R.E.; Page, A.J.; Alikhan, N.; Fornika, D.; Maguire, F.; Mendes, C.I.; Tausch, S.H.; Black, A.; Connor, T.R.; Tyson, G.H.; Aanensen, D.M.; Alcock, B.; Campos, J.; Christoffels, A.; Gonçalves da Silva, A.; Hodcroft, E.; Hsiao, W.W.; Katz, L.S.; Nicholls, S.M.; Oluniyi, P.E.; Olawoye, I.B.; Raphenya, A.R.; Vasconcelos, A.T.R.; Witney, A.A.; MacCannell, D.R. The PHA4GE SARS-CoV-2 Contextual Data Specification for Open Genomic Epidemiology. *Preprints* (2020).

Figures

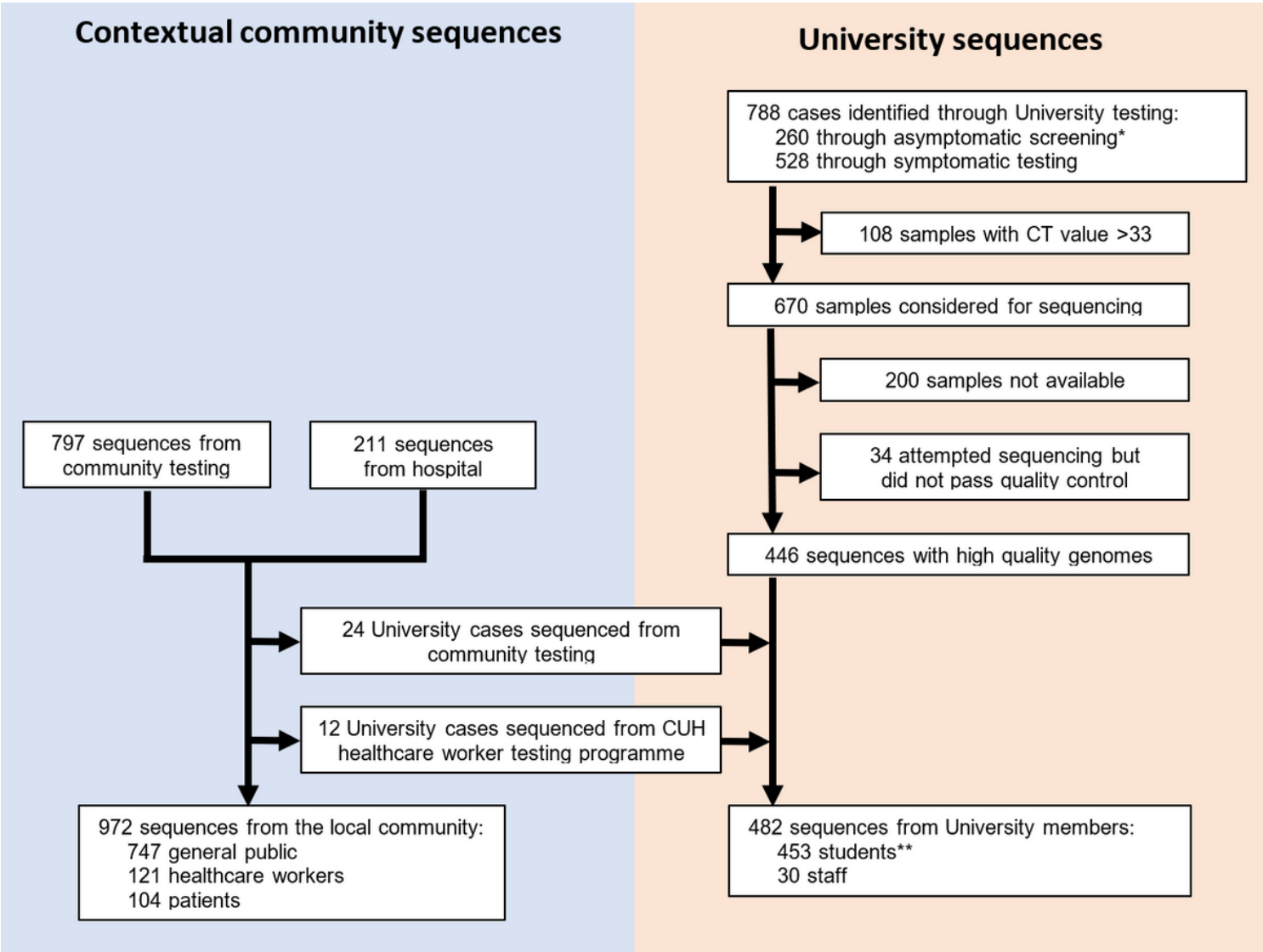

Figure 1

Study cohort and available genome sequences *Includes 14 students identified through ad hoc asymptomatic screening conducted as part of an outbreak investigation by the University of Cambridge in conjunction with local public health authorities, responding to increased rates of infection in a block of student accommodation (described in further detail in cluster 2 below). **includes 2 students associated with a single sequenced pooled sample (see supplementary methods). CUH = Cambridge University Hospitals

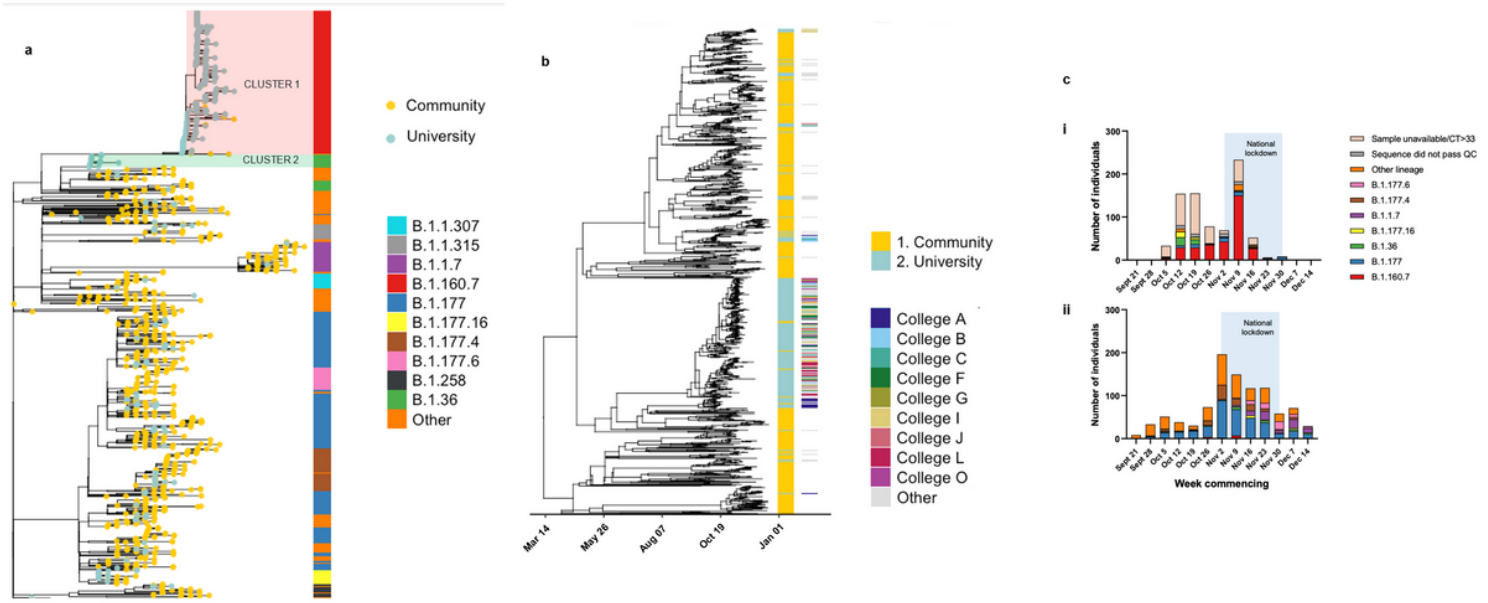

Figure 2

a: Maximum likelihood tree showing that the majority of lineages from university isolates were distinct from community isolates. The node leaves (branch tips) show case location and global PANGO lineage is illustrated in the vertical bar. b: Time-scaled coalescent tree including university members and local community isolates from study period with visible segregation between the two groups. College affiliation is shown for university members in the second set of vertical columns, highlighting the 'top nine' colleges by cluster 1 prevalence. c: Epidemic curves demonstrating a steeper decline in SARS-CoV-2 cases in the University of Cambridge (i) compared to the local community (ii), with associated lineages. Only cases with available genomes are included. University term ran from the week commencing October 5 to the week commencing November 30. The light blue shaded area reflects a 4-week national lockdown in the UK, which was associated with a large fall in COVID-19 cases in University students. Specific lineages highlighted are the 4 largest lineages within the University (minimum 20 cases over the study period) and the community (minimum 50 cases over the study period). For (i), weekly individual case ascertainment for staff and students testing positive for SARS-CoV-2 through both symptomatic and asymptomatic testing pathways provided at the University of Cambridge is indicated. For (ii), weekly cases with genomes available from the local community are shown.

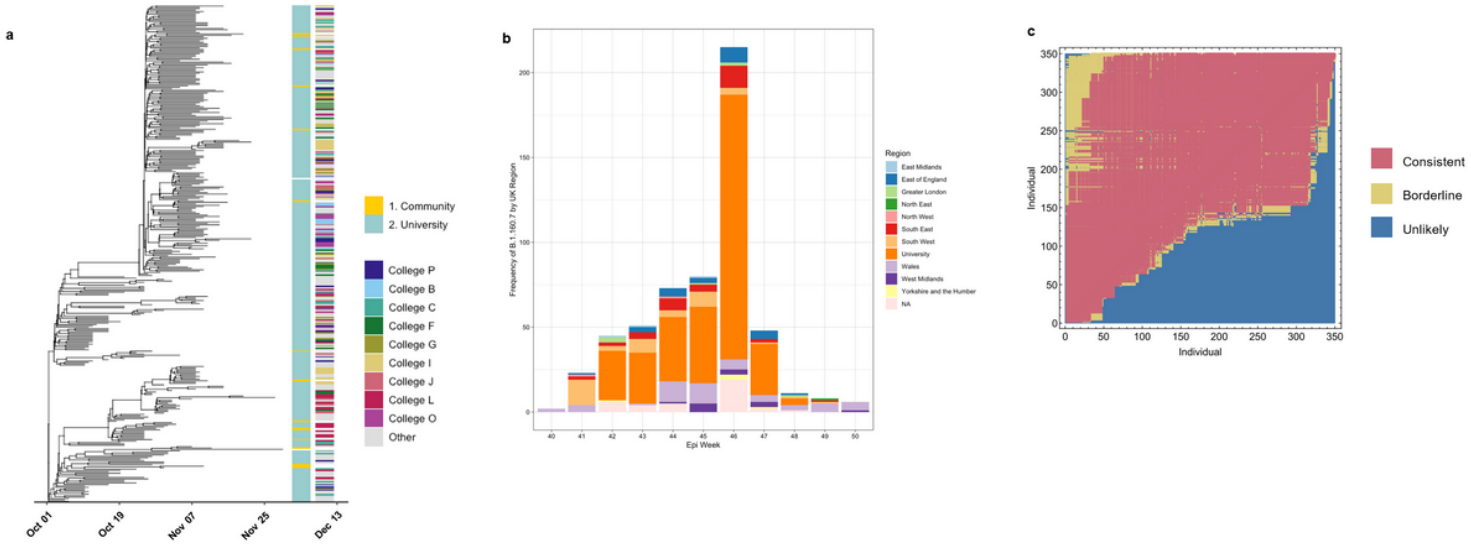

Figure 3

a: Time-scaled phylogenetic tree of largest university cluster (cluster 1) derived from the BDSKY model implemented in BEAST 2.6 (see Figure 5). The left-sided heatmap is coloured by case location, and the right-sided heatmap is coloured by student college affiliation, highlighting the 'top nine' colleges by cluster 1 prevalence. Cluster 1 was widely dispersed across the university with limited transmission into the community. b: Frequency of Lineage B.1.160.7 (to which cluster 1 belongs) in each region of the UK and the University of Cambridge. Regions are defined as 'Nomenclature of territorial units for statistics' (NUTS) regions, where the UK has 9 regions. It is visible that the lineage B.1.160.7 was first sequenced in Wales, and then in the neighbouring South West of England, before becoming prevalent within the University of Cambridge. The lineage remained infrequently detected in the community populating the wider surrounding region (East Anglia, Bedfordshire and Hertfordshire, and Essex, making up East of England) throughout the university term. c: A continuous transmission chain of SARS-CoV-2 infections in cluster 1 commenced with a single introduction. Relationships between individuals in cluster 1 were calculated within A2B-COVID. Colours denote potential transmission events from the donor (vertical axis) to the recipient (horizontal axis) that are consistent with transmission 12 or which are borderline possibilities (yellow). The plot shows that the data are consistent with a continuous transmission chain of SARS-CoV-2 infections in cluster 1 occurring via a single introduction.

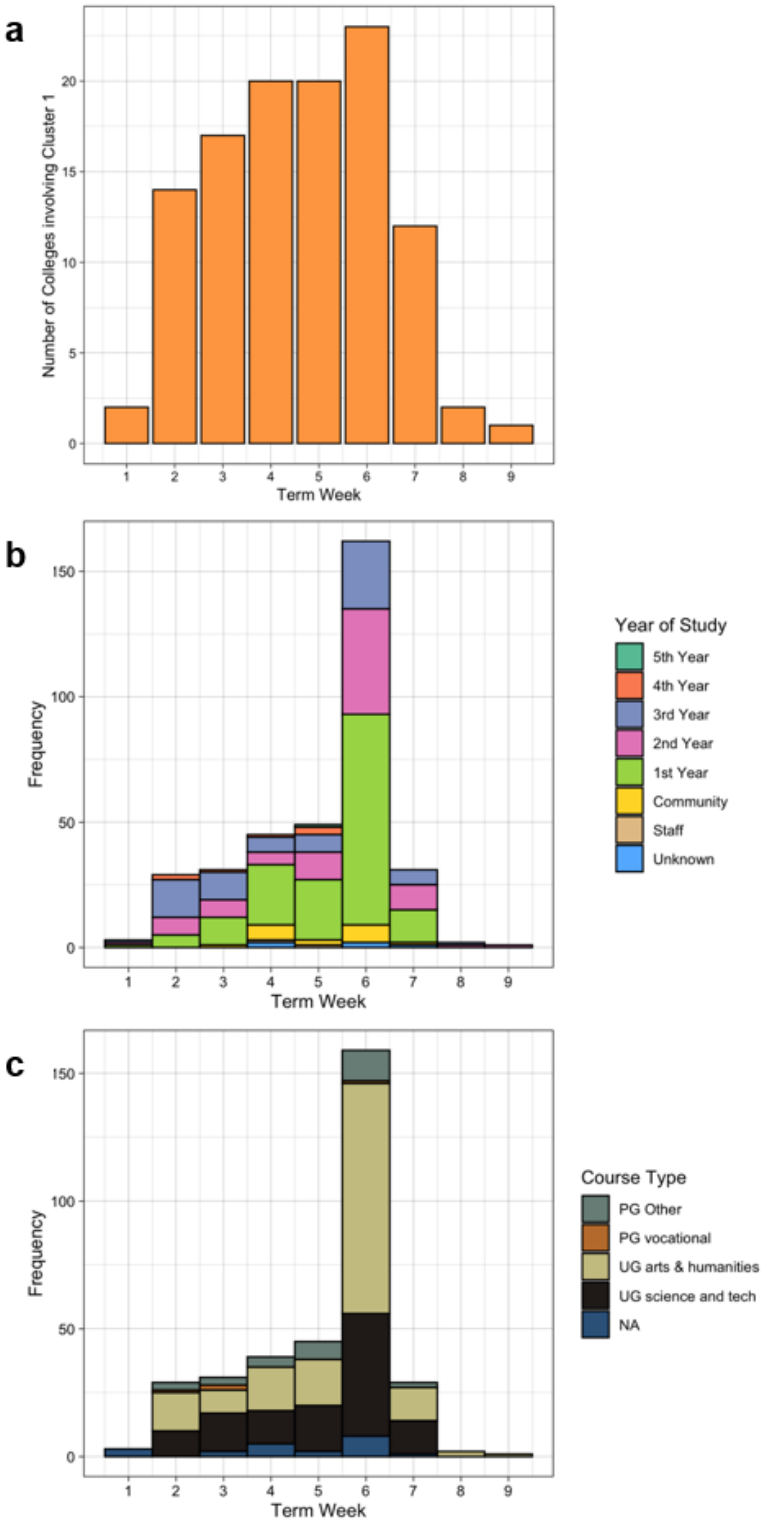

Figure 4

Demographics of Cluster 1 across the 1st university term. A) Cumulative number of colleges involved in the cluster. Cases included in this cluster were between a number of colleges early during the university term. B) Frequency of cases involved in the cluster by course type. C) Frequency of cases involved in the cluster by year of study.

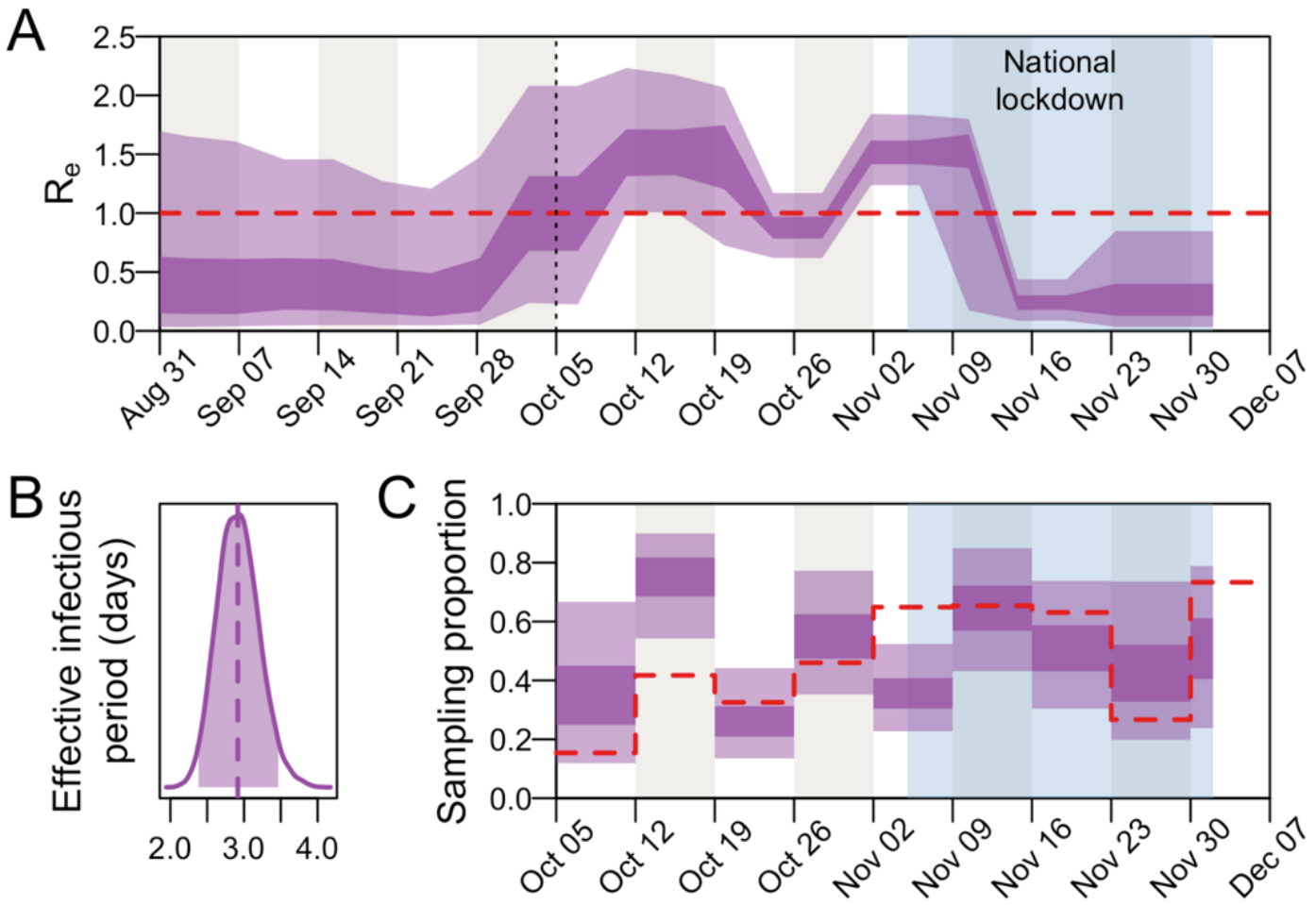

Figure 5

A 20-epoch birth-death skyline model shows the effect of local infection control measures and the national lockdown on the effective reproduction number (Re), and estimates of the mean effective infectious period as 2.9 (95% HPD=2.4-3.5) days. (A) Re posterior estimates (dark shading=50% HPD; light shading=95% HPD). The dotted line indicates the start of term and the light blue shaded area the 4-week national lockdown in the UK, which was associated with a large fall in COVID-19 cases in University students. The red dashed line indicates Re=1. (B) Effective infectious period posterior estimates (shaded region=95% HPD; dashed line=median). (C) Weekly sampling proportion posterior estimates (dark shading=50% HPD; light shading=95% HPD). The red dashed line indicates the empirical sampling proportion estimates for each week in term (number of sequenced genomes from all University clusters divided by the number of positive tests among University staff and students).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ExtendedDataFiguresfinal.docx

- SupplementaryMethodsandResultsfinal.docx
- SupplementaryTablesfinal.docx
- COGUKauthorship01122020Nohighestqualification.docx
- UoCCOVIDScreeningConsortiumandCOVIDTestingConsortium.docx